U.S. Army Research Laboratory

SUMMER RESEARCH TECHNICAL REPORT

Examining a Terrorist Network Using Contingency Table Analysis

ALLISON MOORE
MENTOR: DR. ELIZABETH K. BOWMAN
COMPUTATIONAL AND INFORMATION SCIENCES DIRECTORATE
ABERDEEN PROVING GROUND, MARYLAND

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
AUG 2011

2. REPORT TYPE

3. DATES COVERED
00-00-2011 to 00-00-2011

4. TITLE AND SUBTITLE
Examining A Terrorist Network Using Contingency Table Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Computational And Information Sciences Directorate,Aberdeen Proving Ground,MD,21005

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
See also Ada548876

14. ABSTRACT
The information revolution of the 21st century has changed the nature of war to focus on the area of network-centric warfare. While the number and strength of dark networks continue to increase, the U.S. Department of Defense seeks to identify, predict, and counteract possible terrorist threats. The intelligence community is undertaking the seemingly impossible task of using the information that is currently available through social networks and military reports. This report will examine the Ali Baba data set that was created in 2003 for the National Security Agency (NSA) by Mark Jaworoski and Steve Pavlak. The Ali Baba data set contains fictitious Word documents that have recorded the actions of a suspected terrorist network. This report will demonstrate the use of statistics in examining terrorist organizations. Specifically, it will determine if contingency table analysis using the R programming language can be used to analyze a terrorist network.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**16**	

Contents

List of Figures

List of Tables

Abstract

The information revolution of the 21st century has changed the nature of war to focus on the area of network-centric warfare. While the number and strength of dark networks continue to increase, the U.S. Department of Defense seeks to identify, predict, and counteract possible terrorist threats. The intelligence community is undertaking the seemingly impossible task of using the information that is currently available through social networks and military reports. This report will examine the Ali Baba data set that was created in 2003 for the National Security Agency (NSA) by Mark Jaworoski and Steve Pavlak. The Ali Baba data set contains fictitious Word documents that have recorded the actions of a suspected terrorist network. This report will demonstrate the use of statistics in examining terrorist organizations. Specifically, it will determine if contingency table analysis using the R programming language can be used to analyze a terrorist network.

Acknowledgments

I wish to acknowledge the mentorship of Elizabeth Bowman.

Student Bio

I will be a senior this fall at Virginia Polytechnic Institute and State University in Blacksburg, Virginia. I am a double-major in Mathematics and Statistics, with a minor in Actuarial Science. This is my second year as a summer student at the U.S. Army Research Laboratory (ARL). After graduation, I plan on either attending graduate school to concentrate in applied statistics or becoming a mathematical statistician for the government. Prior to this summer, I had no experience working with the R programming language.

1. Introduction/Background

The information revolution of the 21st century has changed the nature of war (*2*). The areas of cyber-warfare and network centric warfare (*2, 4*) are an expanding focus of research within the United States military and the Department of Defense (DoD) (*7*). While the number and strength of networks continue to increase, the intelligence community is developing programs and techniques to identify, predict, and counteract possible terrorist threats (*3, 4*). The intelligence community is undertaking the seemingly impossible task of using the information that is currently available through social networks and military reports. Faced with an endless supply of data, this is a very difficult problem.

The purpose of this paper is to demonstrate the use of statistics in examining terrorist organizations. Specifically, we wish to determine if Contingency Table Analysis (CTA) using the R programming language can be used to analyze a dark network. While the statistical scope of this paper is limited, it is my desire to see an increase in the use of statistics in solving problems within the intelligence community. The data set that will be used for all subsequent analysis in this paper is entitled "Ali Baba." In 2003, Mark Jaworowski and Steve Pavlak originally created the data set for the National Security Agency (NSA) as a way to test visualization software since there were no reliable data sets available for evaluation. They sought to develop a data set that an analyst would be able to piece together and determine the intentions of a terrorist cell.

The Ali Baba data set contains fictitious word documents that have recorded the actions of a suspected terrorist network from April to September 2003. The documents include local communication intercepts and various intelligence reports (police, human intelligence (HUMINT), Foreign Broadcast Information System (FBIS), and detainee). There are a core group of messages (n=75) that contains the main activities of the cell. The remaining messages contain possibly interesting information from peripheral members. The scenario of the data set follows a suspected terrorist cell of radical Islamic fundamentalists that are centralized in England. The FBIS reports in southern Egypt from spring of 2003 describe an outbreak of cholera among large groups of schoolchildren. The radical Islamic world was blamed for the attack, and this data set reveals their desire to revenge the claim. It is the organization's plan to contaminate London's water supply by blowing up a water treatment facility (*9*).

2. Experiment/Calculations

The Ali Baba data set contains 609 entities with nine original variables, as shown in table 1. In order to perform statistical analysis, it is necessary to categorize the data. Table 2 holds the categorized version of the data set that contains a total of five variables. This specific process of categorization is labor-intensive, as each line has to be assigned by hand due to variability in the data.

Table 1. Original Ali Baba variable table.

Variable Name	Variable Type	Description
Document	Character	Source of message and identification number
Date	Character	Date of original message
Associate1	Character	Author of message
Associate2	Character	Receiver of message or mentioned in message
Associate3	Character	Receiver of message or mentioned in message
Location	Character	Where the message originated
Event/Activity	Character	Summary of message contents
Organization	Character	Organization involvement of the author
Targeted Entity	Character	Target of interest

Table 2. After categorization Ali Baba variable table.

Variable Name	Variable Type	Description
Truth	Numeric	Villain level of Associate 1, 2, 3
Location	Character	Region of the message's origin
Event	Character	Categorization of event/activity
binlocal	Numeric	If Location = "SE" or "London" then binlocal = 1, else = 0
binTruth	Numeric	If Truth = 1 then binTruth = 1, else = 0

The biggest change between table 1 and table 2 is the switch of the "Associate" variables from character to the ordinal numeric "Truth" variable. This is done to denote the members through what we will call Core Villain Levels, which range from 1 to 5. A Ground Truth document created by Jaworoski and Pavlak (*8*) is used to rank the villains Black, Dark Blue, Medium Blue, Light Blue, and Unknown, or "1", "2", "3", "4", and "5," respectively. Two issues arise when categorizing the "Associate" variables. First, if an individual is ranked as a "5," we will say that they are not involved with the terrorist organization, and is, therefore, not a suspected threat. Second, if an individual was a possible member of two core villain levels, they were given the higher ranking. The levels provided by the Ground Truth document could also be determined by a social network graph like those created in Analyst Notebook.

The variable "Location" is categorized to represent the change from individual locations to regions (*11*). Additional entities for "Location" include "Unknown", "Outside" for locations outside of Britain, and "Britain" for unspecific messages. Several obvious errors were found while going through the "Location" variable of the original data set. These errors include London being spelled "Lodnon," or "Brigton" instead of Brighton. These mistakes are corrected to ensure more accurate results. The "Event/Activity" variable is also changed to "Event," which describes the change to categorized groups based on certain keywords found in the data set. For example, the "Brit" grouping is based on messages that mentioned Tony Blair, British government or politics, and anti-British or anti-UK sentiment. This particular grouping is the most controversial, as it is determined particularly by me.

The final two variables in table 2 are binary variables called “binlocal” and “binTruth”. In regression, binary variables also known as indicator variables take on the values “0” or “1” to indicate failure or success of a categorical effect (*1*). The binary results for “Location” are recorded as “1” when the villain sends a message from “SE” or “London,” and recorded as “0” otherwise. Similarly, the “binTruth” is set as a “1” when the villain is a Black Villain, and recorded as “0” otherwise.

3. Results and Discussion

Now that all of the data is categorized, we are able to fully use CTA within the R programming language. CTA can be used to investigate the relationship between two or more variables (*1*).

First we can examine the Villain Truth Level vs. Location table, which is table 3. The primary step is to determine if the two variables are independent or not. We conduct a Pearson’s Chi Square test (*1*) in R, and determine that X^2 = 95.1836, degrees of freedom = 30, and p-value = 1.053 x 10^{-8}. The degrees of freedom are determined by (# of rows – 1)*(# of columns – 1). Since the p-value is less than .05, we can reject the null hypothesis and conclude that Truth and Location are dependent.

$$X^2 = \sum \frac{(n_{ij} + \mu_{ij})^2}{\mu_{ij}} \tag{1}$$

Table 3. Truth vs. location.

	Brit.	E	EM	Lon.	NE	NW	Out	SE	SW	Unk.	WM	Totals
Black	2	4	1	75	2	0	11	34	5	5	1	140
Dark	10	8	0	50	1	0	15	34	15	20	1	154
Medium	4	4	0	56	4	0	3	30	0	9	5	115
Light	1	2	0	40	5	3	6	48	10	2	0	117
Totals	17	18	1	221	12	3	35	146	30	36	7	526

We can also calculate odds and odds ratios from table 4, using formula (*2*).

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{2}$$

Looking at villains overall, we determine that they are 5.3277 times more likely to have sent their message from the Southeast or London than any other location. We will, thus, hypothesize that the terrorist cell is focused in the Southeastern region of Britain. This fact will be useful later when we are interested in creating a model. For all further analysis, we will classify “higher level villains” as Black or Dark Blue Villains, and “lower level” as Medium or Light

Blue. Again, looking at table 3, we can calculate that lower level villains are 4.07 times more likely to send messages from Northern Britain than higher level villains.

Second, we can examine the Villain Truth Level vs. Location table. The Pearson's Chi Square in table 4 is 155.7799 with 27 degrees of freedom, and a calculated p-value of less than 2.2x10^{-16}. Therefore, we can reject the null hypothesis and conclude that the variables are dependent. Various odds and odds ratios can once again be calculated from the table. For example, upper level villains are 8.482 times more likely to be found discussing non-suspicious items, while Light Blue villains are over five times more likely to discuss casing than their counterparts.

Table 4. Truth vs. event.

	Alq	Assoc	Bomb	Brit	Case	Islam	911	Not	Rec	Sus	Totals
Black	*4*	*36*	*16*	*4*	*1*	*11*	*4*	*33*	*20*	*11*	140
Dark	*11*	*33*	*9*	*4*	*3*	*45*	*9*	*21*	*16*	*3*	154
Medium	*6*	*35*	*10*	*7*	*2*	*27*	*4*	*7*	*11*	*6*	115
Light	*1*	*26*	*11*	*4*	*29*	*25*	*2*	*0*	*6*	*13*	117
Totals	22	130	46	19	35	108	19	61	53	33	526

Lastly, we can look at the Location vs. Event table (table 5). The chi-square calculation from this table is 315.5827, with 90 degrees of freedom and a p-value of less than 2.2x10^{-16}. Therefore the "Location" and "Event" variables are dependent on each other. Interesting odds ratios from the table are that villains in the Southeast or London are 3.7796 and 2.1237 times more likely to discuss suspicious things and Al Qaeda, respectively, than villains in other locations. Using the data in the three previous tables, various graphs, figures, and odds ratios can be created. Due to the required brevity of this paper, these additional items have not been included. Through the information gathered in previous tests, we are able to begin the process of developing a model to fit the data.

Table 5. Location vs. event.

	Alq	Assoc	Bomb	Brit	Case	Islam	911	Not	Rec	Sus	Totals
Brit.	*1*	*1*	*11*	*0*	*0*	*0*	*0*	*1*	*1*	*1*	16
E	*6*	*7*	*3*	*6*	*1*	*10*	*5*	*8*	*1*	*3*	50
EM	*0*	*0*	*1*	*1*	*0*	*0*	*2*	*3*	*0*	*0*	7
Lon.	*13*	*35*	*16*	*14*	*3*	*43*	*7*	*29*	*24*	*21*	205
NE	*0*	*6*	*0*	*4*	*0*	*1*	*6*	*1*	*1*	*1*	20
NW	*0*	*0*	*2*	*0*	*0*	*1*	*1*	*0*	*1*	*0*	5
Out	*7*	*7*	*3*	*0*	*0*	*4*	*7*	*3*	*3*	*1*	35
SE	*18*	*25*	*6*	*8*	*23*	*22*	*34*	*20*	*3*	*22*	181
SW	*0*	*1*	*3*	*4*	*4*	*8*	*1*	*8*	*3*	*5*	37
Unk.	*1*	*4*	*1*	*4*	*4*	*10*	*21*	*1*	*0*	*1*	47
WM	*0*	*1*	*4*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	6
Totals	46	87	50	41	35	99	84	74	38	55	609

Using R and the "Design" package, we wish to construct a model (*8*, *12*) from the Ali Baba data set. The data has a two-level variable called "binlocal" that is used as our response or dependent variable. The predictors or independent variables from table 2 are "binTruth" and "Event." Before the model is created, we must check the various tables to see if any cells are small or empty. There are several methods in R that can be used to create easy-to-read tables from an inputted data set, as seen earlier. The two-way tables for binTruth vs. binlocal, binlocal vs. Event, and binTruth vs. Event are satisfactory to continue the analysis. We then wish to determine if the variables in the three tables are independent or dependent. This is done with a chisq.test line of code in R and compared at the 5% significance level. The three variable pairings have a p-value of less than .05, so we can conclude that the respective variables are dependent.

Now that the basic assumptions have been satisfied, we are able to look at our first logistic regression model. Figure 1 contains the partial output from R for the initial model that uses all three of the variables discussed previously.

Looking at the table of coefficients in the output (figure 1), we see that the intercept and several of the factors of "Event" are not significant in the model. A second model (figure 2) is created that removes the "Event" variable to see if we can improve our results.

```
Logistic Regression Model

lrm(formula = binlocal ~ binTruth + Event, data = fullset, na.action =
na.pass)

Frequencies of Responses
  0   1
159 367

       Obs  Max Deriv Model L.R.       d.f.           P
       526      1e-05      56.96         10           0

            Coef    S.E.   Wald Z P
Intercept   -0.3321 0.4359 -0.76  0.4462
binTruth     0.8224 0.2556  3.22  0.0013
Event=assoc  1.3448 0.4818  2.79  0.0052
Event=bomb  -0.4098 0.5342 -0.77  0.4431
Event=brit   0.7158 0.6475  1.11  0.2689
Event=case   2.6822 0.7447  3.60  0.0003
Event=isl    1.1722 0.4841  2.42  0.0155
Event=nin    0.2715 0.6361  0.43  0.6695
Event=not    0.4826 0.5178  0.93  0.3513
Event=rec    1.4077 0.5530  2.55  0.0109
Event=sus    1.6087 0.6289  2.56  0.0105
```

Figure 1. Initial logistic model.

```
Logistic Regression Model

lrm(formula = binlocal ~ binTruth, data = fullset, na.action = na.pass)

Frequencies of Responses
  0   1
159 367

       Obs  Max Deriv Model L.R.       d.f.          P
       526      2e-14       6.15          1     0.0131

          Coef   S.E.   Wald Z P
Intercept 0.7009 0.1081 6.48   0.0000
binTruth  0.5564 0.2305 2.41   0.0158
```

Figure 2. Second logistic model.

The outputs in figures 1 and 2 show logistic regression models that are based on the Ali Baba data set. The frequency and observation sections allow us to check that our data was read into R correctly and that we can continue analysis. If these numbers did not match, we would need to go back to our initial R code to check for hidden errors that would not return error messages on the output screen. From figure 1, we can see that the likelihood ratio chi-square for the full model is 56.96, with 10 degrees of freedom, and a p-value of approximately 0. The model without "Event" has a chi-square value of 6.15, degrees of freedom 1, and p-value of .0131. This tells us that both of our models as a whole are statistically significant as compared to a model with no predictors.

The rest of the results show the table of coefficients, their standard errors, the Wald z-test, and the p-values. In figure 2, we can see that both the intercept and the "binTruth" variables are statistically significant at the 5% level. It is important to note that the output is given in ordered logits since we are using logistic regression. To interpret "binTruth," we can say that for a one-unit increase (going from 0 to 1), we expect a .56 increase in the expected value of "binlocal" on the log odds scale. The "Design" package also allows us to output the coefficients as proportional odds ratios as seen below.

```
Effects              Response : binlocal

 Factor      Low High Diff. Effect S.E. Lower 0.95 Upper
0.95
 binTruth    0    1    1     0.56   0.23 0.10       1.01
  Odds Ratio 0    1    1     1.74     NA 1.11       2.74
```

Figure 3. Proportional odds ratios.

Since "binTruth" is an indicator variable, the low and high will always be 0 and 1. For "binTruth," going from 0 to 1 increases the odds of the message location being in the Southeast or London by 1.74 times, given that the other variables are held constant. It is important to note that the proportional odds assumption does hold for the coefficients in this regression.

4. Summary and Conclusions

The process of categorization was the first necessary step to examine the Ali Baba data set using CTA. Once that was completed, we were able to determine if the variables were related with any type of dependence. After this analysis, the location of the messages was centralized and various statistical odds were provided, as well. Further CTA statistics were generated but were not included due to the brevity of this paper. Finally, initial models were created using R packages, but additional research will be conducted this summer on the topic.

With the massive amounts of information available today, there is an increasing need to understand and interpret data provided by social networks and military reports. The modeling and predictive power of statistics in examining dark networks is currently underused. The use of CTA on this data set was time-consuming because of the layout of the data. However, a data set with additional entries, variables, or categories would better have exploited the abilities of CTA. With respect to R, there are various other statistical packages on the market that can provide equivalent results. However, it was the purpose of this paper to focus specifically on the R programming language. The R programming language is preferred by some because of the open-source nature of the language and for the availability of help resources. Overall, the R programming language has proved effective in analyzing the Ali Baba data set in the realm of CTA.

5. References

1. Agresti, Alan. *An Introduction to Categorical Data Analysis.* 2^{nd} ed. Hoboken, NJ: John Wiley & Sons, Inc., 2007.

2. Alberts, David S.; Garstka, John J., Stein, and Frederick P. *Network Centric Warfare.* Center for Advanced Concepts and Technology, 1999. Print.

3. Allanach, Jeffrey; Tu, Haiying; Singh, Satnam; Willett, Peter; Pattipati, Krishna. Detecting, Tracking, and Counteracting Terrorist Networks via Hidden Markov Models. University of Connecticut: Dept. of Electrical and Computer Engineering, n.d. Web. 22 June 2011. www.engr.uconn.edu/~sas03013/docs/Aerospace_HMM.pdf.

4. Arquilla, John; Ronfeldt, David. The Advent of Netwar. *Networks and Netwars: The Future of Terror, Crime, and Militancy*. Rand, 19 Mar. 2002. Web. 22 June 2011. http://faculty.cbpp.uaa.alaska.edu/afgjp/padm610/networks%20and%20netwar.pdf.

5. Boik, John. Lab 4: Correlation, Regression, Contingency Tables. *R Computing*. 30 Nov. 2007. Stanford University. Web. 13 June 2011. www.stat.stanford.edu/~jcboik/stat-141-2007/R_labs/07Lab4.pdf.

6. "Entering Data."*R Class Notes*. UCLA: Academic Technology Services, Statistical Consulting Group. Web. 9 June 2011. www.ats.ucla.edu/stat/r/notes/entering.htm.

7. Flynn, Michael T.; Pottinger, Matt; Batchelor, Paul D. Fixing Intel: A Blueprint for Making Intelligence Relevant in Afghanistan." *Center for a New American Security*. 04 Jan. 2010. Web. 12 July 2011. http://www.cnas.org/node/3927.

8. Harrell, Jr. Frank E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis.* Springer, New York, 2001.

9. Jaworowski, Mark; Pavlak, Steve. Ali Baba Scenario 1 Description. 2003. Print. 1 June 2011.

10. Kabacoff, Robert I. Frequencies and Crosstabs. *Quick-R: Accessing the Power of R. Web. 10 June 2011*. www.statmethods.net/stats/frequencies.html.

11. Map of England's Regions. Outdoor Sport & Leisure. n.d. Web. 21 June 2011. www.outdoor-sport-leisure.net/activities-UK.htm.

12. Ordinal Logistic Regression. *R Data Analysis Examples*. UCLA: Academic Technology Services, Statistical Consulting Group. Web. 11 July 2011. www.ats.ucla.edu/stat/r/dae/ologit.htm.

13. Quinn, Kevin. Log Linear Example." 11 Dec. 2002. University of Washington. Web. 13 June 2011. www.stat.washington.edu/quinn/classes/536/S/loglinexample.html.